Published online 7 December 2011

Nucleic Acids Research, 2012, Vol. 40, No. 4 ***e30***
doi:10.1093/nar/gkr1205

POLYPHEMUS: R package for comparative analysis of RNA polymerase II ChIP-seq profiles by non-linear normalization

Marco A. Mendoza-Parra*, Martial Sankar, Mannu Walia and Hinrich Gronemeyer*

Department of Cancer Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC)/CNRS/INSERM/Université de Strasbourg, BP 10142, 67404 Illkirch Cedex, France

Received August 4, 2011; Revised November 17, 2011; Accepted November 18, 2011

ABSTRACT

Chromatin immunoprecipitation coupled with massive parallel sequencing (ChIP-seq) is increasingly used to map protein–chromatin interactions at global scale. The comparison of ChIP-seq profiles for RNA polymerase II (PolII) established in different biological contexts, such as specific developmental stages or specific time-points during cell differentiation, provides not only information about the presence/accumulation of PolII at transcription start sites (TSSs) but also about functional features of transcription, including PolII stalling, pausing and transcript elongation. However, annotation and normalization tools for comparative studies of multiple samples are currently missing. Here, we describe the R-package POLYPHEMUS, which integrates TSS annotation with PolII enrichment over TSSs and coding regions, and normalizes signal intensity profiles. Thereby POLYPHEMUS facilitates to extract information about global PolII action to reveal changes in the functional state of genes. We validated POLYPHEMUS using a kinetic study on retinoic acid-induced differentiation and a publicly available data set from a comparative PolII ChIP-seq profiling in *Caenorhabditis elegans*. We demonstrate that POLYPHEMUS corrects the data sets by normalizing for technical variation between samples and reveal the potential of the algorithm in comparing multiple data sets to infer features of transcription regulation from dynamic PolII binding profiles.

INTRODUCTION

Based on a technological leap in the development of sequencing technologies, we are currently facing a switch from gene centric to global analyses. However, the concomitant development of bioinformatics tools that allow for comparative functional analyses is lagging significantly behind, such that a lot of presently available data have not yet been exploited to gain maximal functional insight. Chromatin immunoprecipitation coupled with massive parallel sequencing (ChIP-seq) is one of these technologies, which is increasingly used to define biologically relevant processes like (the dynamics of) chromatin modifications and constituents at genome-wide scale, the association patterns of (post-translationally altered) chromatin modifiers or chromatin interacting proteins, such as transcription factors or RNA polymerases. Multiple computational approaches dedicated to ChIP-seq have been developed to (i) generate signal intensity profiles by cumulating sequenced reads aligned to the genome and (ii) identify significant 'peaks' over the reconstructed profile (1).

For RNA polymerase II (PolII) the corresponding ChIP-seq profile is the composite of at least two functionally different aspects; a strong and well-defined binding to a given transcription start site (TSS) where PolII is observed even in absence of transcriptional activity (2) and a second composite pattern resulting from PolII action subsequent to transcription initiation at a given TSS, which comprises several regulated events like transcript elongation, pausing and termination (Supplementary Figure S1). Therefore, any comparative study of transcriptional activities at distinct gene loci should consider the global behaviour of read-count signal intensities spread over the entire loci and not rely

*To whom correspondence should be addressed. Tel: +33 3 88 65 34 19; Fax: +33 3 88 65 34 37; Email: marco@igbmc.fr
Correspondence may also be addressed to Hinrich Gronemeyer. Tel: +33 3 88 65 34 73; Fax: +33 3 88 65 34 37; E-mail: hg@igbmc.fr
Present address:
Martial Sankar, Department of Plant Molecular Biology, University of Lausanne, Biophore Building, CH-1015 Lausanne, Switzerland.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

only on the comparison of PolII occupancies at TSSs. Moreover, such a comparison should provide a genome-wide read-out that helps to gain insight into the regulated functional aspects of PolII subsequent to the initiation of transcription. Finally, variations inherent to the technology (i.e. sequencing depth) should be considered in the comparison of different data sets by applying adequate normalization procedures as for the differential expression analysis of mRNA-Seq assays (3).

Here, we describe POLYPHEMUS, an R package that integrates ChIP-seq derived PolII binding site information with gene annotation to identify coding regions associated with transcriptional activity and compare changes of these activities in biological contexts. For this, POLYPHEMUS performs gene length standardization and read-counts intensity of non-linear normalization of multiple data sets before comparison with correct technical/experimental variations, which could cause problems for comparative data analysis. We show here through the analysis of primary data and meta-analyses of published data sets that POLYPHEMUS can be used as an integral part of an analytical pipeline (Supplementary Figure S2) for the comparative analysis of multiple PolII ChIP-seq data sets to gain functional insight about the differential activity of gene networks activities in different biological contexts.

MATERIALS AND METHODS

Overview

POLYPHEMUS is based on an integrative approach which combines, in a user-directed manner, information from several levels, as is schematically illustrated in Supplementary Figure S3. (i) As the first step POLYPHEMUS combines peak calling outputs from PolII ChIP-seq experiments with coding region database annotations; from this information, signal intensity profiles are extracted, which reveal those coding regions at which PolII was identified by ChIP-seq. (ii) The signal intensity profiles are scanned with a sliding window, thus producing smoothened sliding window intensity profiles. (iii) To compare two distinct PolII ChIP-seq experiments both their sliding window intensity profiles and (iv) the coding region lengths are normalized. (v) Finally, normalized profiles are displayed as a relative differential enrichment for PolII association. The procedures developed for peak calling/coding region data integration, normalization and standardization are described in detail below.

Identification of RNA PolII-enriched coding regions

As a first step, we identify chromatin sites to which PolII is bound. For this, we use MeDiChI, a model-based deconvolution approach originally developed by Reiss *et al.* (4) to study ChIP-ChIP profiles and which has been adapted for ChIP-seq data analysis (5). Whereas other peak caller outputs can be used together with POLYPHEMUS, MeDiChI provided an efficient manner to annotate significant PolII-enriched regions thanks to a peak-shape learning process that is performed before annotation of enriched-regions [illustrated in Supplementary Figure S4 in comparison with the widely used peak caller 'Model-based analysis of ChIP-Seq (MACS)' (6)].

Subsequent to PolII binding site identification POLYPHEMUS correlates peak positions with a coding region annotation database for the organism of interest, such as RefSeq (7). For this, the genomic locations of the identified PolII peaks are compared with annotated Transcription Start Sites (TSS) within a user defined window (default ± 300 bp) around peak centres. The overlap identifies coding regions for which the ChIP-seq analysis displays significant enrichment of PolII at the TSSs. Together with the signal intensity wiggle files, this information is used to extract read-count intensities along the corresponding coding regions. To smoothen the PolII ChIP-seq profile over the gene bodies, a user-defined sliding window (default 250 bp) scans the concerned coding regions to compute a median sliding-window intensity (SWI). User-defined buffer regions (default 500 bp) upstream and downstream of the concerned genes are included in the analysis to include ChIP-seq-defined PolII binding that extends beyond annotated coding regions. Finally, the orientation of genes encoded by the negative strand is inversed to facilitate the comparative analyses in the subsequent steps.

Normalization of RNA PolII profiles

Before comparing the signal intensities within ChIP-seq data sets, it is essential to know if their global amplitudes are indeed comparable. Considering that the amplitude of ChIP-seq profiles is directly proportional to the total number of mappable reads (TMRs), previous studies have normalized different samples by linear correction with a scaling factor that adjusts for TMRs between samples (8–12) (Table 1), following the assumption that the differences in the TMRs uniformly affect the amplitude of the profile. To assess whether this assumption is valid, we displayed the SWI distribution pattern of

Table 1. Normalization approaches used for the analysis of RNA Polymerase II ChIP-seq profiles

Normalization method	Approach	Software implementation	References
Sequencing depth	TMRs uniformly equalized relative to the sample with the lower number of reads	No	(9,10)
Linear scaling	The average reads count in a defined bin is divided by the TMRs	No	(7,8,11)
LOWESS	Locally weighted polynomial least square regression applied to estimate the mean and variance between the compared data sets	No	(13)
Quantile/LOWESS	Described in this study	POLYPHEMUS (R package)	This manuscript

compared profiles as minus versus average (MA) transformation plot, which is frequently used in microarray data analysis (13). Importantly, we observed that the differences in the TMRs can result in rather dramatic non-linear deviation of the compared SWIs (for example, see Figure 1B top and bottom panels), indicating that a reliable comparison of ChIP-seq data sets with different TMRs could require in certain cases more sophisticated procedures than linear scaling.

This issue has been addressed recently by applying locally weighted polynomial least square regression (LOWESS) to estimate the smoother line of the mean and the variance of the observed data (14). POLYPHEMUS has LOWESS functionality integrated, but to include the possibility of comparing multiple profiles, we implemented, in addition, a quantile normalization option. The rationale for this is that while LOWESS and quantile normalizations produce similar results, there are two limitations when using LOWESS. First, span conditions to obtain the best smoothening (proportion of points used to compute) need to be empirically evaluated, thus making automation impossible and second, LOWESS requires high computation time, which is a serious disadvantage when dealing with next-generation sequencing data in a genome-wide context. Note that the implementation of LOWESS normalization in POLYPHEMUS follows a similar procedure as described (14).

Quantile normalization. Quantile normalization relies on the assumption that the majority of coding regions present the same transcriptional activity across the compared experimental conditions, which reflects a common PolII association pattern to constitutively active genes. Correspondingly, the quantile normalization adjusts the distribution of SWIs for different samples to reach a common distribution pattern (15), by the following procedure:

For N ChIP-seq data sets with n PolII-enriched coding regions, each of which comprising Z sliding windows:

(i) build a matrix M of size $K \times N$, where each column is a ChIP-seq data set (N) and where each row correspond to the SWIs per coding region. Note that the total number of rows are defined by $K = \sum_{i=1}^{n} Z_i$
(ii) sort each column of M to give M_{sort}
(iii) take the means across each row of M_{sort} and scale (i.e. divide) each SWIs with this value to get M'_{sort}
(iv) get the final matrix M_{norm} by rearranging each colum of M'_{sort} to have the same ordering than M

In contrast to LOWESS, this approach can be extended to more than two samples, which, for instance, allows for the comparison of samples from kinetic analyses. The current POLYPHEMUS version handles multisample normalization.

Comparing normalized ChIP-seq profiles. Subsequent to normalization, compared profiles are expressed by the ratios of their corresponding normalized SWIs. To correct for variations between contiguous SWI ratios, we fit the distribution of each coding region-specific ratios with a LOWESS-smoothened line. Each ratio is then interpolated to the fitted line, defined as fitted SWI ratios.

Defining TSS/gene body regions. Different types of transcription regulation-relevant information can be extracted

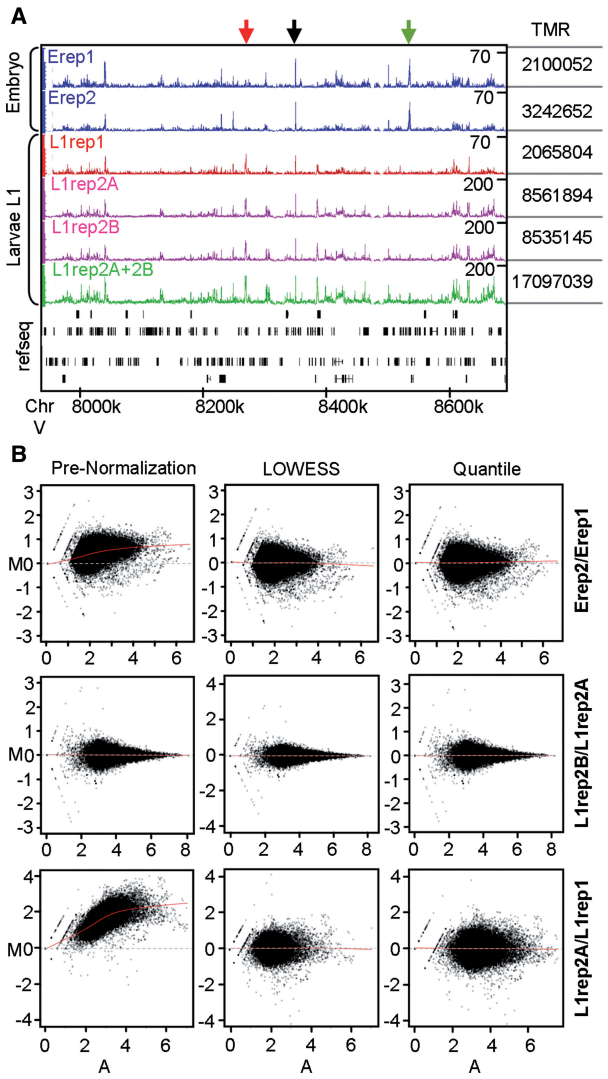


Figure 1. Comparison of RNA PolII ChIP-seq profiles requires non-linear normalization. Meta analysis of *C. elegans* PolII profiling by ChIP-seq (10). (**A**) The signal tracks for chromosome V illustrate the different samples which are compared in this study. Display of two biological replicates (suffix 'rep1' or 'rep2') of samples from embryo (E; blue tracing) or L1 larval (L1; red and pink tracings) stages. For the biological replicate 2 of the L1 sample two technical replicates (L1rep2A and L1rep2B) are displayed. TMR for these samples range from 2–8 million. Upregulated (red arrow), constitutive (black arrow) and downregulated (green arrow) PolII binding at given loci can be intuitively detected by visual inspection of the profiles, as different signal intensity scales are displayed (1:200 for the L1 high TMR sample and 1:70 for samples with about 2 million TMR). (**B**) MA plots. The fitted LOWESS curve (red line) in the prenormalized MA plot for Erep2 versus Erep1 (top left) reveals the need for nonlinear normalization before data comparison. LOWESS (top center) or quantile normalization (top right) was applied to enable data comparison. MA plots for the technical replicate (L1rep2B versus L1rep2A) and a biological replicate (L1rep2A versus L1rep1) are also illustrated before or after LOWESS/quantile normalization.

from PolII association at TSS and gene body regions when studying gene regulation. Therefore, POLYPHEMUS has been designed such that binding at the TSS and the pattern observed along the coding region of a gene provide separate readouts. This approach facilitates gene classification according to local PolII activities, such as stalling or productive elongation (illustrated in Figures 3A and B and 5A and C). Inspired by previous studies describing the PolII enrichment around TSS sites of annotated genes (16–18), we have defined a 250-bp distance from the TSS as being indicative of relevant PolII-TSS binding (user-defined parameter in POLYPHEMUS). The remaining part of the coding region is defined as 'gene body'.

Gene length standardization. To compensate for the highly variable length of gene coding regions POLYPHEMUS normalizes gene lengths. Consequently, an equal number of data points define all coding regions, which is an essential prerequisite for comparative analyses. The procedure is performed as following: for a given gene body composed of Z sliding windows, where Z_j corresponds to their positions in the coding region of interest, the body-length standardization to a reference length L (in sliding window units) is performed by the transformation

$$l_j = \frac{L}{Z} z_j$$

where l_j correspond to the standardized positions of the sliding windows in the gene-body. As each sliding window presents a given SWI, such information is represented in the context of their standardized positions (l_j) and fitted to a LOWESS-smoothened line, which is then used to interpolate the number of data points that will represent the body gene characteristics for PolII binding.

Classification of PolII-occupied coding regions

POLYPHEMUS (i) combines identified PolII binding sites with signal intensity profiles, (ii) normalizes sample data sets for subsequent comparison and (iii) standardizes different coding regions such that comparative intracoding region and intersample analyses become possible. Given the existence of efficient tools for data clustering, we did not integrate such option into POLYPHEMUS. Rather, POLYPHEMUS generates a versatile matrix output (in addition to MA plots and intensity tables), in which columns correspond to normalized and standardized sliding window intensity (SSWI) ratios and lanes to the corresponding coding regions (Supplementary Figure S5). This matrix can be uploaded in tools like MultiExperiment Viewer (MeV) (19,20) to perform supervised or unsupervised gene clustering analysis based on normalized SWI ratios covering the corresponding coding regions, as exemplified below and depicted in Figures 2, 3 and 5.

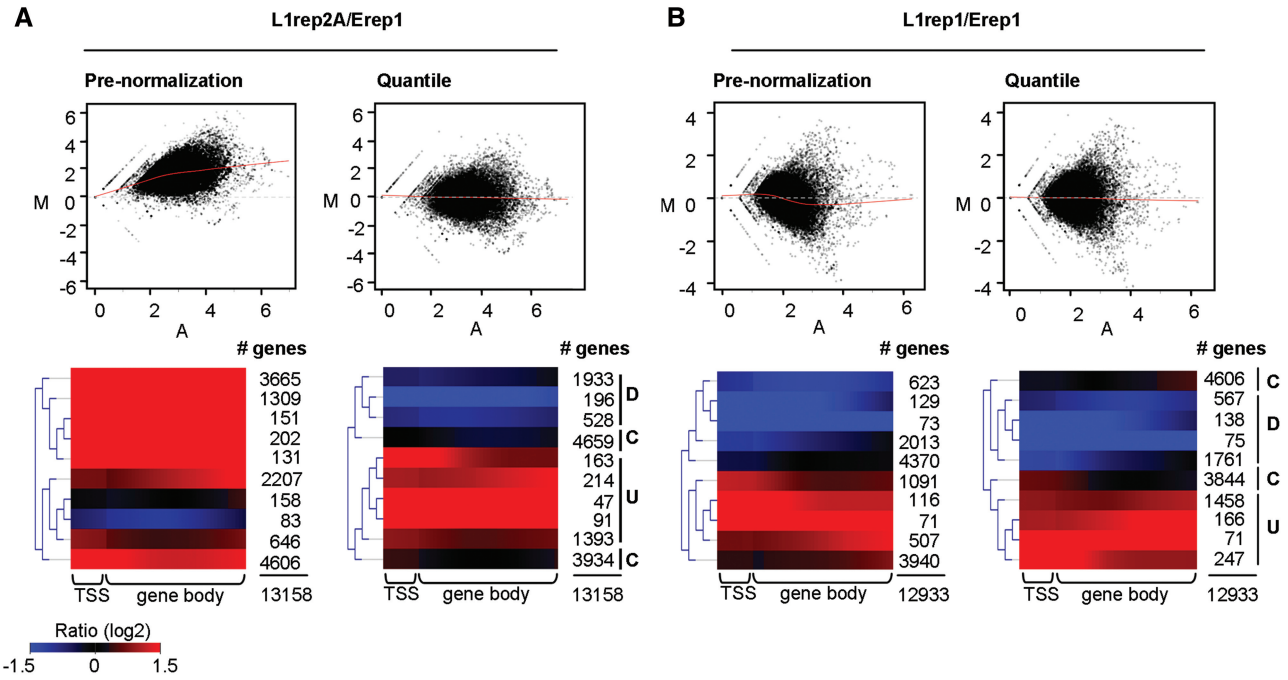


Figure 2. Analysis of differential chromatin-association of RNA PolII at different developmental stages requires data normalization. Comparison of ChIP-seq data sets for larval and embryonic stages for (**A**) high TMR difference (L1rep2A versus Erep1: 8.5 versus 2.1 million reads) and (**B**) similar TMR (L1rep1 versus Erep1: 2.0 versus 2.1 million reads). (Top panels) MA plots are shown before ('Prenormalization') and after normalization ('quantile'). (Bottom panels) SOTA (max cycles = 9; cell variability, $P = 0.01$) to classify PolII associated genes according to the ratios of the signal intensities for each annotated gene in the two compared samples; shown is the SOTA before and after quantile normalization. The different SOTA-predicted classes are catalogued according to their relative PolII binding characteristics: constitutive (C: ~65%), downregulated (D: ~20%) and upregulated (U: ~15%).

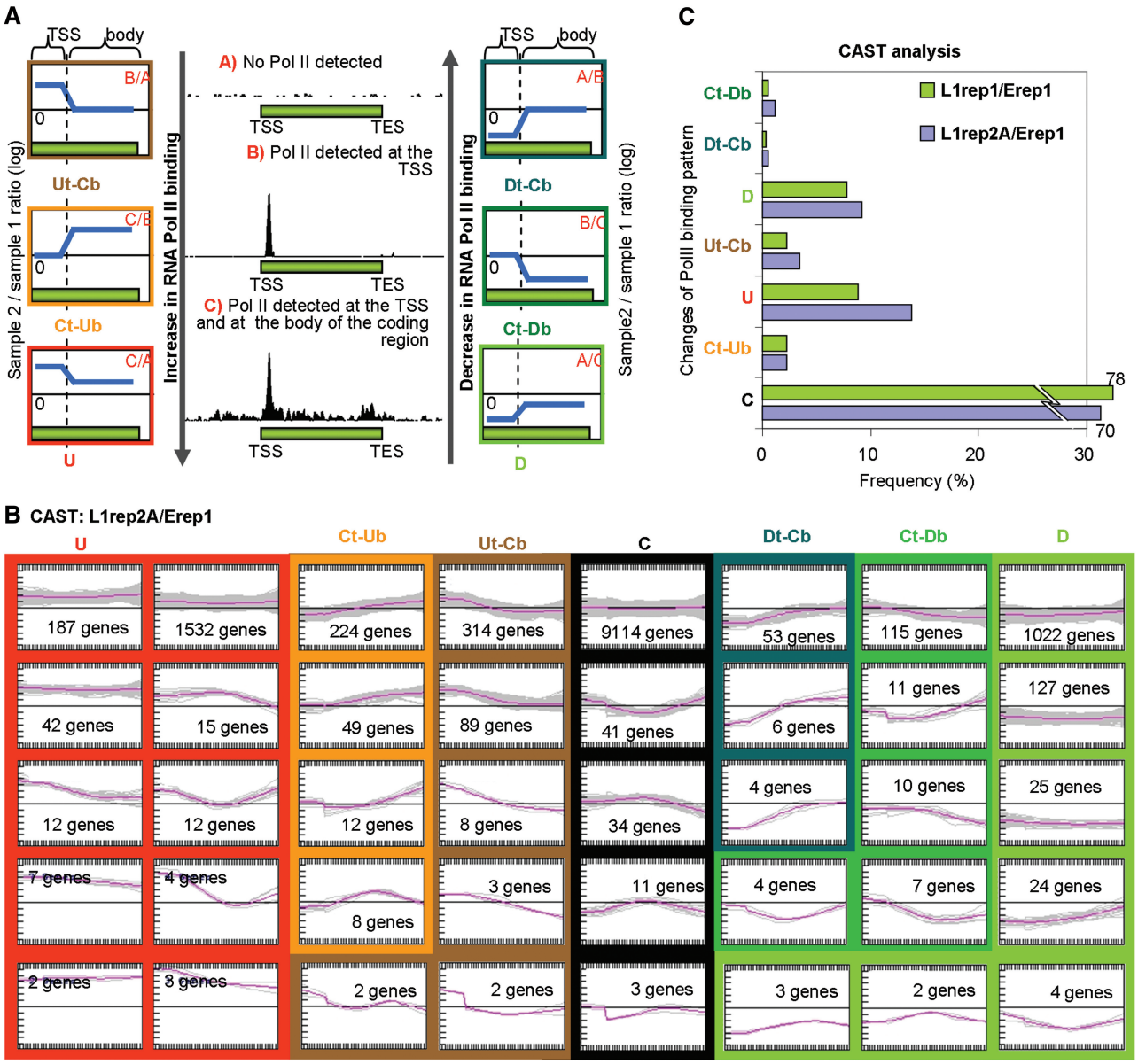


Figure 3. Sub-classification of RNA PolII binding by TSS-gene body profiling. (**A**) Schematic representation of the conceptual PolII binding patterns in a comparative analysis (Ut-Cb upregulated signal intensity at TSS, constitutive gene body; Ct-Ub, constitutive TSS and increased signal intensity at gene body; U, global increase of the signal intensity; Dt-Cb, decrease in signal intensity at the TSS and constitutive pattern at gene body; Ct-Db, signal intensity constitutive at TSS and decreased at gene body; D, global decrease of signal intensity). (**B**) Illustration of the different PolII binding patterns from comparing L1rep2A with Erep1. CAST (similarity cutoff: 0.9) has been used to perform the initial classification, followed by an intuitive association of classes into those depicted in (a). (**C**) Frequency of the different transcriptional binding patterns indicated in (a) from comparing L1rep1 or L1rep2A with Erep1.

RESULTS

Meta analysis of *Caenorhabditis elegans* RNA PolII chromatin association characteristics at different developmental stages

To test and validate our approach, we have applied POLYPHEMUS to a publicly available *C. elegans* (GEO accession number: GSE15628) data set that contains ChIP-seq profiles for PolII binding performed in the context of a study of PHA-4/FOXA transcription factor binding in embryonic and starved L1 larval stages (12). We chose this data set, as a plethora of controls are provided for each PolII ChIP-seq sample. Specifically, the authors have generated two biological replicates for embryo and L1 larval stages. In addition, two technical replicates are reported for one of the L1 biological replicates. Hereafter, these different RNA PolII ChIP-seq samples are referred as Embryo-rep1 (Erep1), Embryo-rep2 (Erep2), Larvae L1-rep1 (L1rep1) and L1-rep2A (L1rep2A) as well as L1-rep2B (L1rep2B; Figure 1A). In addition, a signal intensity profile obtained by compiling the mappable reads in L2A and L2B (L1rep2A + B) is included for comparison (Figure 1A).

Non-parametric normalization methods implemented for proper comparison of RNA PolII ChIP-seq binding profiles

As expected, the comparison of PolII signal intensity profiles obtained for larval L1 and embryo stages reveals different patterns (Figure 1A). Indeed, gains and losses of PolII occupancy (red and green arrow, respectively) signify differential transcriptional activities at the associated coding region. Although visual inspection of the illustrated ChIP-seq profile supports this interpretation, a global comparison of all data points clearly reveals the need for data normalization. Indeed, the TMR differences between samples required a display of the wiggle intensity profiles at different scales to allow for a visual inspection (Figure 1A; compare L1rep2A: ~8 million, scale 200, and L1rep1: ~2 million reads, scale 70). Obviously, such an adjustment bears the risk that in a comparison of samples apparent differences for the observed peak intensities of a given PolII binding site are due to technical variability (TMRs differences) rather than the consequence of regulation.

While such comparative analysis strongly suggest the necessity of data normalization, only an analytical approach, like the MA transformation of the profile-associated SWIs may reveal whether a linear correction or a more sophisticated approach suffices for correcting the described differences. As shown in Figure 1B, even biological replicates with small TMR differences can exhibit an important offset behaviour in their MA plots towards high signal intensity values; this aberration increases with increasing TMR difference (Figure 1B; compare top panel Erep2 versus Erep1 with a difference of ~1 million reads and bottom panel L1rep2A versus L1rep1 with a difference of ~6 million for the TMR). Note that the MA plots of the technical replicates L1rep2B and L1rep2A (ΔTMR: 26 749 reads) reveal a well-centred pattern relative to the x-axis. The above analysis reveals that in this particular situation, a linear scaling approach is not suitable for normalization when comparing samples, even if TMR differences are as low as 1 million. For this reason, we have implemented LOWESS and quantile normalization as user-defined options in POLYPHEMUS to generate cohorts with comparable data distribution. (Figure 1B middle and right panels).

Monitoring differential chromatin association of RNA PolII at different developmental stages

The main interest in comparing signal intensity levels of PolII tracings is to infer differences in the transcriptional features of related biological materials. Zhong *et al*. (12) compared two different developmental stages in *C. elegans*, larval L1 and embryo stages, to assess chromatin localization of the transcription factor PHA-4/FOXA and correlate its locus-specific binding with the regulation of gene expression. For this, they pooled the PolII ChIP-seq data sets from biological replicates, followed by linear TMR-based normalization of signal intensity profiles. To evaluate the need for linear or non-linear data normalization, when comparing such samples from different developmental stages, we have performed MA plots for high and low TMR differences (Figure 2; L1rep2A versus Erep1, ΔTMR ~6 million; L1rep1 versus Erep1, ΔTMR ~34 000). A major divergence of data scatters towards high signal intensities is obvious for the high ΔTMR samples (Figure 2A, top left) but even at low ΔTMR, the MA plots reveal a slight offset at both low and high signal intensities (Figure 2B, top left). Both these aberrations were efficiently corrected by quantile normalization (Figure 2A and B, top right). Given that L1rep2A and L1rep1 are biological replicates, a similar differential PolII binding pattern at the TSS and along the gene bodies would be expected when comparing them separately to the embryo data. To validate this assumption, we applied the Self-Organizing Tree Algorithm [SOTA; Euclidean distance; max cycles = 9; cell variability $P = 0.01$ (21)] in MeV (20) to classify PolII binding patterns with or without prior normalization by quantile. Without normalization, the comparison between L1rep2A and Erep1 would indicate a globally increased PolII binding to the majority of loci (Figure 2A, bottom left); this is a consequence of the overall higher signal amplitudes of the L1rep2A data set (Figure 2A, top left). After quantile normalization, the SOTA analysis visualizes the differential PolII binding patterns in both cases. Indeed, around 2500 genes (~20% of the analysed genes) revealed a downregulation of PolII binding, whereas 2000 genes (~15% of the analysed genes) showed an increase when the embryo turns into a L1 larvae. Most notably, comparing Erep1 with either of the biological replicates L1rep1 or L1rep2 yielded similar results after quantile normalization. Importantly, linear correction in case of high TMR differences (L1rep2A versus Erep1) produces aberrant readouts as illustrated in Supplementary Figure S6A.

Classifying RNA PolII binding characteristics at coding regions

The association of PolII with genes follows very complex patterns, which reveal aspects of its chromatin association and processivity; for instance, PolII may get bound to promoters and remain stalled, it may engage in transcription at low rate with high promoter occupancy or all promoter-bound PolII may transcribe the corresponding gene leaving an 'empty' promoter behind. Intermediates between these extremes are likely to exist, given that various effectors, such as PolII-modifying enzymes, factors involved in elongation or ncRNAs regulate PolII-mediated transcription (22).

PolII ChIP-seq profiles are assembled snapshots from a large number of cells visualizing the regulated and dynamic chromatin interaction of enzymes; for each cell, they derive from one or more gene-specific PolII functions that comprise events like promoter loading/TSS occupancy, PolII stalling or travelling along the gene during active transcription or transcription termination. Specific features of a promoter/gene, such as bidirectionality, may have additional impact on PolII binding characteristics. Monitoring the patterns, dynamics and extent of PolII association and correlating these data with PolII function will reveal a readout of genome-wide PolII

transcriptional activity in a given experimental or cell physiological setting.

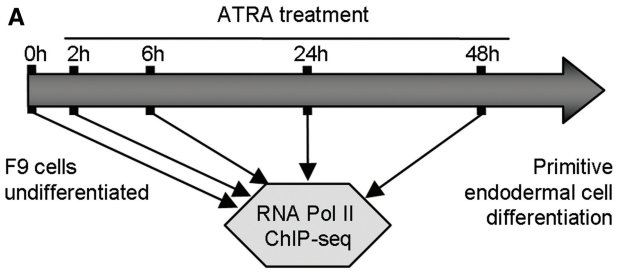
Conceptually, the separate monitoring of PolII binding at the TSS and the gene body may display one of three ChIP-seq profiles for any given gene (Figure 3A): (i) TSS and coding region are deprived of RNA PolII, (ii) only the TSS is occupied and (iii) both TSS and gene body are occupied, or any intermediate thereof. Applied to a comparative situation, like induced gene activation, patterning the ChIP-seq profiles will differentiate between (i) situations where only the TSS shows higher occupancy in one of the data sets but no differences over the coding region ('Ut-Cb' for upregulated TSS-constitutive body); (ii) only the gene body that displays higher occupancy in one data set and no changes are seen at the TSS ('Ct-Ub' for constitutive TSS-upregulated body); (iii) both the TSS and the body show higher PolII occupancy in one data set ('U' for upregulated); (iv) the TSS is deprived of PolII in one data set without any changes over the coding region ('Dt-Cb' for downregulated TSS-constitutive body); (v) the body shows higher occupancy in one data set and no differential behaviour at the TSS ('Ct-Db' for constitutive TSS-cownregulated body); and finally (vi) both the TSS and the body are deprived of PolII in one data set ('D' for downregulated).

While some of the above scenarios can be revealed by SOTA analysis (Figure 2A and B, bottom panels), a classification approach that does not predefine the number of classes can lead to a more refined set of PolII association patterns. To this end, we used the Cluster Affinity Search Technique [CAST; Euclidean distance; threshold affinity value = 0.9 (23)] that is implemented in MeV (20). Figure 3B illustrates the CAST analysis for the comparison between L1rep2A and Erep1 after quantile normalization. This unsupervised clustering generated some 40 classes that can be intuitively reorganized in the above-described conceptual patterns. The upregulated and downregulated PolII binding events previously revealed by SOTA analysis are now retrieved as additional subclasses (U: ~9–13%; Ut-Cb: ~2–3%; Ct-Ub: ~2%; D: ~8–9%; Dt-Cb: ~0.5%; Ct-Db:~1% in Figure 3C). Note that applying CAST separately to the two biological replicates revealed essentially the same classification, as previously demonstrated using SOTA.

RNA PolII binding characteristics during F9 cell differentiation

That POLYPHEMUS is designed to compare multiple data sets makes it the method of choice to analyse temporal PolII binding kinetics at a genome-wide level. To this aim, we used the well-characterized retinoid-induced F9 mouse embryonal carcinoma (EC) cell differentiation model [reviewed in ref. (24)]. Samples were collected during the first 48 h of all-trans retinoic acid (ATRA) treatment and processed for ChIP-seq analysis of PolII binding (Figure 4A; Supplementary File S1 for details). Alignment of the sequenced reads against the mouse genome (mm9) yielded 4–6 million TMR for all samples (Figure 4B). As expected, MA plots for all time points relative to vehicle presented variable degrees of offset behaviour, revealing the need for normalization; the corresponding multicomparison quantile normalization is depicted in Figure 4C (bottom panel).

We used supervised clustering to classify genes into different patterns of relative PolII binding during cell differentiation. The corresponding SOTA [Euclidean distance; max cycles = 9; cell variability $P = 0.01$ (21)], reveals important changes during differentiation (Figure 5A). The relative abundance of the various PolII chromatin binding classes changes rapidly during differentiation, revealing a highly dynamic recruitment/dissociation and/or processivity of PolII (Figure 5B). Two hours after ATRA treatment approximately 900 genes show increased PolII binding (P-value confidence: 0.05; Supplementary Figure S7). At 6 h, the proportion of genes with higher PolII binding pattern is subdivided in two groups, the upregulated group or 'U' characterized by significant



B

Time frame	TMRs
0h	4582361
2h	6171659
6h	5091977
24h	4244928
48h	5389908

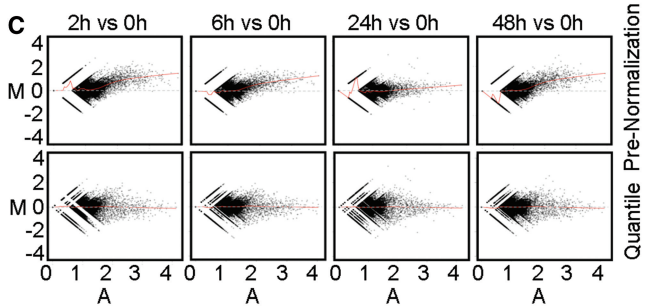


Figure 4. Analysis of the kinetics of RNA PolII binding during differentiation (**A**). Illustration of differentiation system. F9 teratocarcinoma cells were treated with ATRA for 48 h to induce primitive endodermal differentiation. Samples were collected at 0, 2, 6, 24 and 48 h and ChIPed with anti-PolII antibodies for ChIP-seq analyses. (**B**) Table illustrating the TMR per data set. (**C**) MA plots of data sets before and after quantile normalization. All samples were normalized relative to the 0 h control sample. Note that, as in the case of the *C. elegans* data sets (Figures 2 and 3) all data sets require non-linear normalization, as is obvious from the LOWESS fitted line (red) in the prenormalization MA plots.

PolII ratio levels at the TSS as well as the gene body (558 genes) and that characterized by significant PolII ratio levels preferentially at the TSS (Ut-Cb; 529 genes). At 24 h, the number of genes revealing upregulated RNA PolII binding decrease dramatically (approximately 200 genes are classified as 'U' and approximately 700 genes as 'Ut-Cb,' respectively). However, this trend is reversed 48 h after ATRA treatment. At this time, more than 2500 genes display significant PolII levels (relative to 0 h) preferentially localized at the TSS (Ut-Cb). This biphasic PolII recruitment pattern correlates with two transcription peaks previously reported for F9 cell differentiation (25). Note that in contrast to PolII recruitment, the number of genes that lose PolII binding (relative to the 0 h) remained rather constant, suggesting that the majority of loci are 'poised' for transcription activation by early or constitutive recruitment of PolII.

Notably, SOTA classification not only identifies recruitment, loss and constitutive binding of PollI, but it shows binding patterns that diverge between TSS and the body of the coding regions, as illustrated for *Nanog*, *Stra8* and *Cdv3* (Figure 5C). The PolII ChIP-seq signal intensity

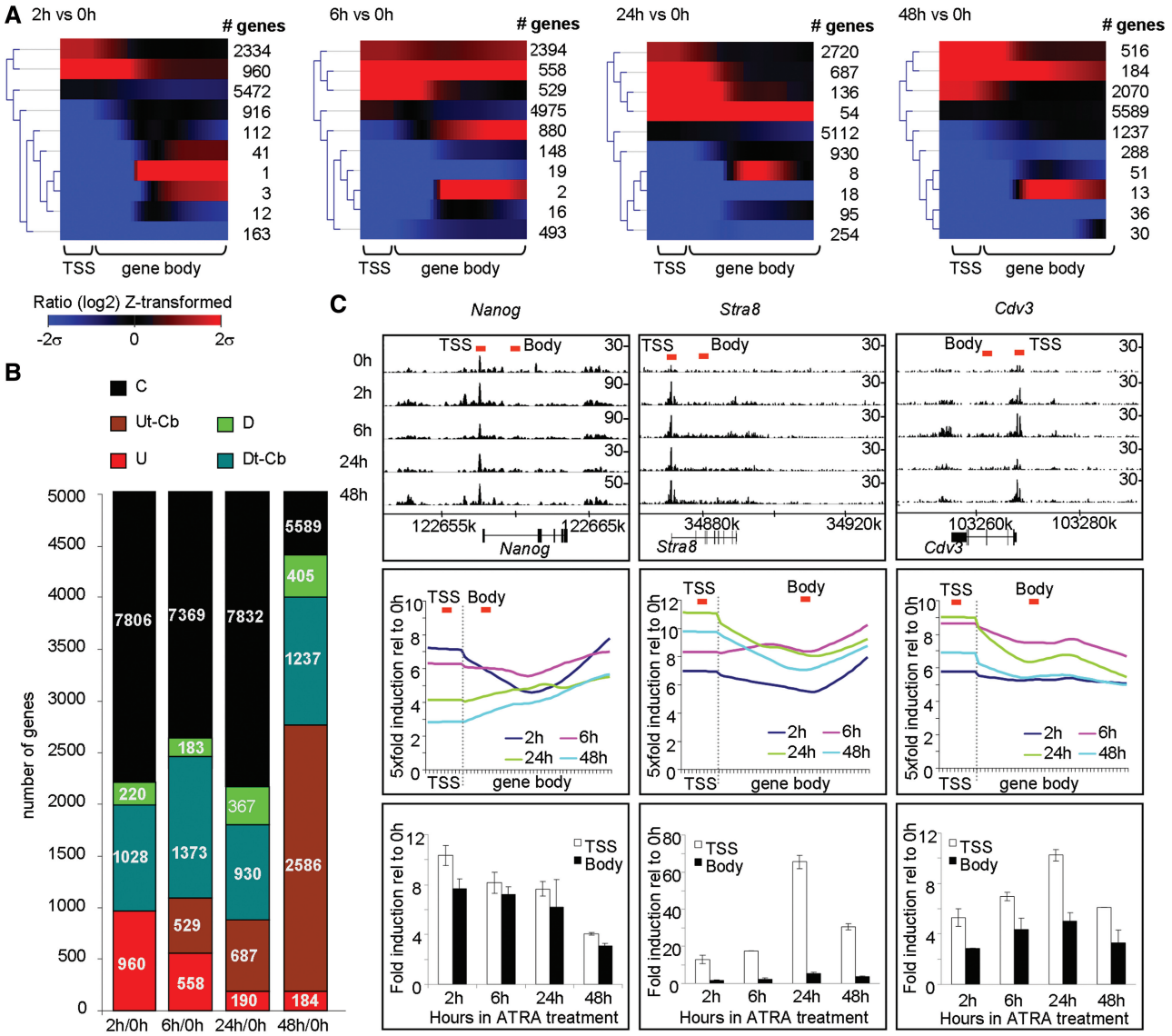


Figure 5. Dynamic chromatin association patterns of RNA PolII during ATRA-induced differentiation. (**A**) POLYPHEMUS (QUANTILE)-normalized ATRA-treated data sets were classified relative to the nontreated control using SOTA (max cycles = 9; cell variability $P = 0.01$). The associated heat maps to the SOTA analyses illustrate the presence of upregulated (red), constitutive (black) and downregulated (green) PolII binding at identified loci based on at least 2σ distance away from the global mean behaviour (Supplementary Figure S7). (**B**) The SOTA analysis of (A) was classified according to the occupancy of TSS, gene body or both and the number of coding regions per classes is depicted. The nomenclature is as in Figure 3. (**C**) Three examples illustrate the highly dynamic chromatin association of PolII during F9 cell differentiation. The signal intensity profiles (top panel) are compared with the normalized gene representation (middle panel). The corresponding qPCR validation performed at the TSS and at a defined region of the coding sequence is depicted in the bottom panel. The POLYPHEMUS representation (middle panel) displays the coding region in the *X*-axis (the TSS and the Body regions are delimited) and the fold induction of PolII binding levels at a given time point of ATRA treatment relative to the non-treated control.

profiles demonstrate that, whereas the presence of transcriptional activity on such regions is evident, the differences between time points are rather difficult to assess (Figure 5C, top panels). With POLYPHEMUS, the relative differences between time points become readily apparent (Figure 5C, middle panels). Using quantitative real-time PCR (qPCR) of ChIPed PolII binding sites at the TSS or a position inside the corresponding coding region reveals a good correlation with the computational analysis by POLYPHEMUS (Figure 5C, bottom panels). For example, the PolII levels present at the coding region of *Nanog* appeared to be strongly induced in the first 2 h of ATRA treatment and steadily decreasing thereafter. This aspect was revealed both by ChIP-qPCR analysis and POLYPHEMUS, while it is not evident from the original signal intensity profiles. Similarly, ChIP-qPCR for *Stra8* and *Cdv3* correlated with POLYPHEMUS computation revealing a maximal recruitment of PolII to the TSSs at 24 h of ATRA treatment. Additional features of PolII association along gene loci are apparent from the analysis and warrant further attention for mechanistic analyses, such as a progressive increase (*Nanog*, 24 and 48 h), decrease (*Cdv3*, 6 and 24 h) or U-shaped (*Nanog*, 2 h; *Stra8*, 2, 24 and 48 h in contrast to the 6 h pattern) binding of PolII along the coding region. Such features are likely revealing functional aspects of (regulated) PolII binding to chromatin and/or features of (regulated) transcription in the context of chromatin structure and regulatory machineries.

DISCUSSION

ChIP sequencing is the current method of choice to define and compare genome-wide chromatin binding patterns of regulatory factors, enzymes, non-coding RNAs and chromatin-modifying/transcriptional machineries, as well as posttranslational modifications of chromatin constituents and coregulatory factors, irrespective of whether they bind directly or indirectly to chromatin. In the majority of cases, different data sets serve as a way to identify chromatin regions occupied by several components, but the relative signal intensities associated to the chromatin regions of interest are in most of the cases neglected. As signal intensity profiles are generated from the number of reads that overlap in a given window, there is a direct correlation between the TMR and the amplitude of the signal intensity profile. Previous studies addressing the influence of sequencing depth on the accuracy of binding site annotation showed that the number of identified sites increase with increasing TMR and provided evidence of factor-specific saturation levels. For PolII, saturation at promoter regions is attained beyond 3 million TMR, while it is not reached even at 20 million TMR for the transcription factor STAT1 (26).

One of the most exciting features of ChIP-seq analyses is the possibility to compare different conditions linked to a particular biological/mechanistic question, such as the dynamics of transcription factor binding to its cognate targets during a biological process (cell differentiation, oncogenic transformation, developmental processes, etc.). The problem in such a comparison is the variability of the technique itself. We demonstrate this issue in a meta-analysis of an extensive data set from *C. elegans* in which two different developmental stages were compared (12). While the technical replicates confirmed the high reproducibility of the sequencing technology, a comparison between biological replicates revealed that even rather small TMR differences can lead to significant non-linear offset behaviour of the corresponding data sets in a comparative MA plot, emphasizing the need for data normalization before comparison. We show that linear normalization inadequately addresses the problem, while the non-parametric normalization procedure integrated in the POLYPHEMUS package reliably correct for the aberrant data scattering. Moreover, we demonstrate that the dynamic chromatin association of PolII during cell differentiation can be accurately monitored after data normalization with POLYPHEMUS.

In addition to defining global patterns of binding sites, ChIP-seq profiles contain a wealth of additional information. In particular for PolII, the ChIP-seq signal intensity profiles originate from a plethora of regulatory inputs at the TSS and the travelling of the enzyme along the entire transcription unit. Indeed, both transcription initiation and transcript elongation are highly regulated events (22,27,28). At the TSS PolII recruitment, chromatin modification and structure, preinitiation complex formation and a multitude of other regulatory events control phenomena like PolII binding, stalling or promoter escape, while events such as post-translational modification, association of elongation factors and non-coding RNAs, regulate the travelling and transcript production. ChIP-seq profiles provide readouts for several of these phenomena by revealing among others aspects like promoter clearance, PolII pausing or dissociation. POLYPHEMUS facilitates such analyses, as it generates and visualizes normalized ChIP-seq signal intensity profiles at the TSS and along the gene body as exemplified in Figure 5C (middle panels). Indeed, we were surprised by the gene-specific variability of these profiles, which are likely to reflect the regulatory events affecting PolII–chromatin interaction in a dynamic, gene- and cell-specific manner. We are aware of the possibility that some functional aspects of PolII action may readout in the ChIP profiles through indirect effects; in this respect, PolII mobility (PolII stalling versus travelling, travelling speed, integration in chromatin-associated complexes, etc.) may alter the efficiency by which it is crosslinked to a given locus.

As for normalization of gene expression data obtained by microarray technologies, quantile normalization by POLYPHEMUS relies on the assumption of a common RNA PolII-binding distribution at the majority of genes investigated, which is maintained across the compared experimental conditions (29). Interestingly, a recent study applied quantile normalization for correcting differences in TMR profiles associated with RNA-seq assays and demonstrated an improved decrease in the bias of monitoring differential gene expression relative to qRT-PCR 'gold standard' measurements when compared with other linear correction approaches (3).

With the rapid development of technologies that provide dramatically increasing sequencing depths, protein–chromatin interaction studies will evolve to integrate quantitative aspects of binding/enrichment along the genome. This study illustrates the necessary steps for such analyses in the case of RNA PolII. The general concepts underlying POLYPHEMUS can be extrapolated to other chromatin interactor studies, such as histone modification profiling. Furthermore, the POYPHEMUS pipeline can be combined with other computational efforts like those focused on promoter identification (18,30). Future versions may include additional statistical normalization methods that requiring less assumptions concerning data distribution, like ANOVA-type models (31), with the aim of expanding its use to other protein–chromatin interactors.

AVAILABILITY

POLYPHEMUS is currently available at http://igbmc.fr/Gronemeyer_Polyphemus and on the CRAN network (http://cran.r-project.org/web/packages/polyphemus/). A Bioconductor-compliant package of POLYPHEMUS is being assembled and will be available in spring 2012.

ACCESSSION NUMBER

ChIP-seq data for the temporal characterisation of RNA Polymerase II binding on the F9 model system (5) has been deposited in GEO under accession number GSE30539.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–7, Supplementary File 1.

ACKNOWLEDGEMENTS

We would like to thank Bernard Jost and Serge Vicaire for sequencing library preparation and Solexa sequencing, Stephanie Le Gras for computational Illumina pipeline treatment and all the members of the Gronemeyer laboratory for discussion related to the applications of POLYPHEMUS.

FUNDING

This work was supported by the European Community (LSHC-CT-2005-518417 'EPITRON' LSHM-CT-2005-018652 'CRESCENDO', HEALTH-F4-2009-221952 'ATLAS' and LSHG-CT-2005-018882 'X-TRA-NET'), the Institut National du Cancer (INCa), the Ligue National Contre le Cancer (laboratoire labelisé) and a fellowship of the Association de Recherche Contre le Cancer and the Fondation pour la Recherche Médicale (to M.W.). Funding for open access charge: Ligue National Contre le Cancer.

Conflict of interest statement. None declared.

REFERENCES

1. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
2. Zeitlinger,J., Stark,A., Kellis,M., Hong,J.W., Nechaev,S., Adelman,K., Levine,M. and Young,R.A. (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat. Genet.*, **39**, 1512–1516.
3. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.
4. Reiss,D.J., Facciotti,M.T. and Baliga,N.S. (2008) Model-based deconvolution of genome-wide DNA binding. *Bioinformatics*, **24**, 396–403.
5. Mendoza-Parra,M.A., Walia,M., Sankar,M. and Gronemeyer,H. (2011) Dissecting the retinoid-induced differentiation of F9 embryonal stem cells by integrative genomics. *Mol. Syst. Biol.*, **7**, 538, October 2011 (doi:10.1038/msb.2011.73; epub ahead of print).
6. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
7. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
8. Feng,W., Liu,Y., Wu,J., Nephew,K.P., Huang,T.H. and Li,L. (2008) A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. *BMC Genomics*, **9(Suppl. 2)**, S23.
9. Lefrancois,P., Euskirchen,G.M., Auerbach,R.K., Rozowsky,J., Gibson,T., Yellman,C.M., Gerstein,M. and Snyder,M. (2009) Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, **10**, 37.
10. Nielsen,R., Pedersen,T.A., Hagenbeek,D., Moulos,P., Siersbaek,R., Megens,E., Denissov,S., Borgesen,M., Francoijs,K.J., Mandrup,S. *et al.* (2008) Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev.*, **22**, 2953–2967.
11. Welboren,W.J., van Driel,M.A., Janssen-Megens,E.M., van Heeringen,S.J., Sweep,F.C., Span,P.N. and Stunnenberg,H.G. (2009) ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.*, **28**, 1418–1428.
12. Zhong,M., Niu,W., Lu,Z.J., Sarov,M., Murray,J.I., Janette,J., Raha,D., Sheaffer,K.L., Lam,H.Y., Preston,E. *et al.* (2010) Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet.*, **6**, e1000848.
13. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
14. Taslim,C., Wu,J., Yan,P., Singer,G., Parvin,J., Huang,T., Lin,S. and Huang,K. (2009) Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics*, **25**, 2334–2340.
15. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
16. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
17. Gilchrist,D.A., Dos Santos,G., Fargo,D.C., Xie,B., Gao,Y., Li,L. and Adelman,K. (2010) Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, **143**, 540–551.

18. Sun,H., Wu,J., Wickramasinghe,P., Pal,S., Gupta,R., Bhattacharyya,A., Agosto-Perez,F.J., Showe,L.C., Huang,T.H. and Davuluri,R.V. (2011) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res*, **39**, 190–201.
19. Saeed,A.I., Bhagabati,N.K., Braisted,J.C., Liang,W., Sharov,V., Howe,E.A., Li,J., Thiagarajan,M., White,J.A. and Quackenbush,J. (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134–193.
20. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.
21. Herrero,J. and Dopazo,J. (2002) Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. *J. Proteome Res.*, **1**, 467–470.
22. Selth,L.A., Sigurdsson,S. and Svejstrup,J.Q. (2010) Transcript Elongation by RNA Polymerase II. *Annu. Rev. Biochem.*, **79**, 271–293.
23. Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.
24. Alonso,A., Breuer,B., Steuer,B. and Fischer,J. (1991) The F9-EC cell line as a model for the analysis of differentiation. *Int. J. Dev. Biol.*, **35**, 389–397.
25. Harris,T.M. and Childs,G. (2002) Global gene expression patterns during differentiation of F9 embryonal carcinoma cells into parietal endoderm. *Funct. Integr. Genomics*, **2**, 105–119.
26. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
27. Buratowski,S. (2009) Progression through the RNA polymerase II CTD cycle. *Mol. Cell*, **36**, 541–546.
28. Sikorski,T.W. and Buratowski,S. (2009) The basal initiation machinery: beyond the general transcription factors. *Curr. Opin. Cell Biol.*, **21**, 344–351.
29. Do,J.H. and Choi,D.K. (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol. Cells*, **22**, 254–261.
30. Gupta,R., Wikramasinghe,P., Bhattacharyya,A., Perez,F.A., Pal,S. and Davuluri,R.V. (2010) Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics*, **11(Suppl. 1)**, S65.
31. Xu,J. and Cui,X. (2008) Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. *Bioinformatics*, **24**, 1056–1062.